Protein Sequence Motif

AMOP, a protein module alternatively spliced in cancer cells

Francesca D. Ciccarelli, Tobias Doerks and Peer Bork

This article describes a new extracellular domain – AMOP, for adhesion-associated domain in MUC4 and other proteins. This domain occurs in putative cell adhesion molecules and in some splice variants of MUC4. MUC4 splice variants are overexpressed in several tumours; in particular, they are highly expressed in pancreatic carcinomas but not in normal pancreas. The presence of AMOP in cell adhesion molecules could be indicative of a role for this domain in adhesion.

Mucins are glycosylated proteins involved in lubrication and protection of epithelial cells, as well as their renewal and differentiation [1,2]. At least ten different human apomucin genes have been characterized so far. The transcripts of all these genes share tandemly repeated amino acid regions that are used to bind O-glycans [3]. For a recent classification of all the mucin genes, see Ref. [4]. The *MUC4* gene is expressed in epithelial tissues of different organs. *MUC4* is not expressed in normal pancreas, although it has been reported to be overexpressed in colon carcinomas [5] and pancreatic adenocarcinomas [6,7]. The full-length precursor of MUC4 (sv0-MUC4) is a 2169 residue, membrane-associated protein with a multidomain organization (Fig. 1a). A putative cleavage site follows the von Willebrand type D (VWD) domain, dividing the precursor into the soluble glycoprotein MUC4α and a transmembrane MUC4β growth factor-like subunit. To predict functional features of MUC4, we tried to identify additional domains in the sequence. We thus studied all the sv0-MUC4 regions not covered by any domain annotated in the protein module resource SMART [8,9].

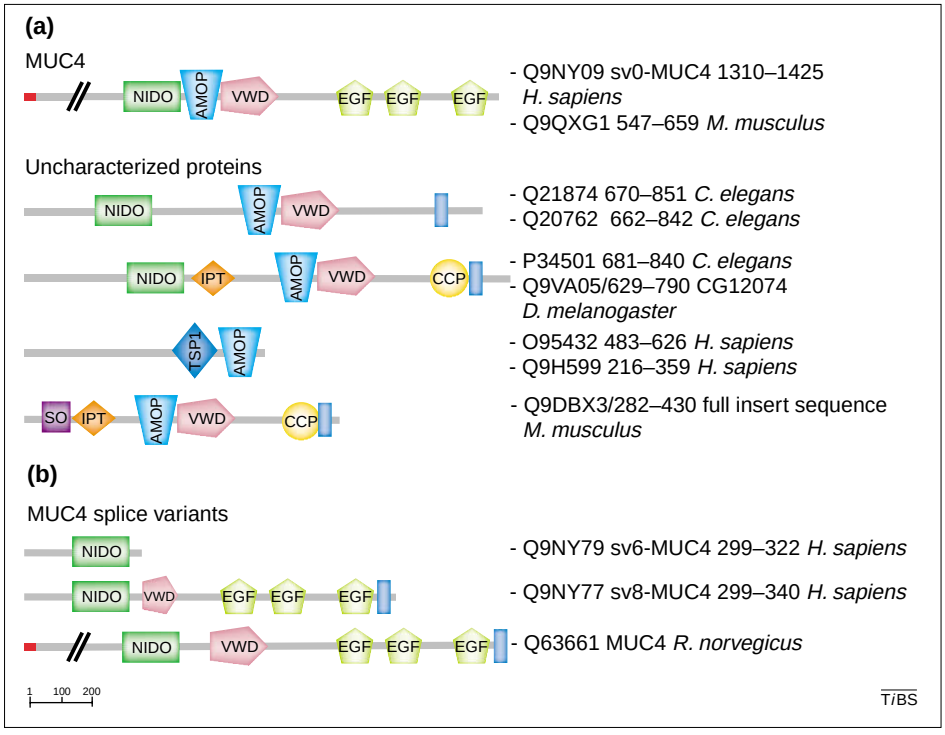

Fig. 1. (a) Domain architectures of AMOP-containing proteins. Only proteins with <80% sequence identity are shown. The domains are named according to the SMART database (http://smart.embl-heidelberg.de) [8,9]. (b) Mucin splice variants showing a deletion of the AMOP domain. Note that in the cases of the MUC4 splice variants Q9NY79 (sv6-MUC4) and Q9NY77 (sv8-MUC4) only a partial conservation of the AMOP domain is present. In particular, in sv6-MUC4 a frameshift occurs after 23 amino acids, and in sv8-MUC4 only the first 41 residues of our domain and the last 31 residues of the VWD domain are conserved. In rat MUC4 the cysteine pattern is not conserved. Blue boxes indicate transmembrane regions. Abbreviations: AMOP, adhesion-associated domain in MUC4 and other proteins; CCP, complement control protein domain, also known as SUSHI repeat or short-complement-like repeat (SCR); EGF, epidermal growth factor domain; IPT, immunoglobulin-like, plexin, transcription factor domain; NIDO, extracellular domain of unknown function, found in nidogen and hypothetical proteins; SO, somatomedin B-like domain; TSP1, type 1 repeat in thrombospondin-1 (binds and activates TGF-β); VWD, Von Willebrand factor type D domain.

Sequence analysis

As PSI–BLAST searches [10] retrieved MUC4 splice variants but not distant homologues for all these regions, we searched for all related proteins with similar domain architectures. One hypothetical protein (accession number P34501) from *Caenorhabditis elegans* also contained a NIDO domain (an extracellular domain of unknown function, found in nidogen and hypothetical proteins) followed by an unannotated region and a VWD domain. The regions between the NIDO and the VWD domains showed a characteristic pattern of cysteine residues in both the *C. elegans* protein and MUC4. We thus performed further PSI–BLAST searches against a non-redundant protein database using the *C. elegans* protein (residues 681–840) as a query, retrieving other proteins with a similar cysteine-rich region (Fig. 2). Although this region frequently precedes VWD domains, it can also be found in other domain contexts (Fig. 1a).

To exclude the possibility that proteins with the novel domain but without a VWD domain showed annotation artefacts, we performed GeneWise searches (http://www.sanger.ac.uk/Software/Wise2/), using the Hidden Markov Model profile of the VWD domain, against the genomic sequences succeeding the coding regions for O95432 and Q9H599. In neither case did we retrieve any region sharing significant sequence similarity to the VWD domain (data not shown). We therefore believe that the newly identified region is a separate domain, and have named this domain AMOP (for adhesion-associated domain in MUC4 and other proteins). It should be

 0968-0004/02/$ – see front matter PII: S0968-0004(01)02049-7

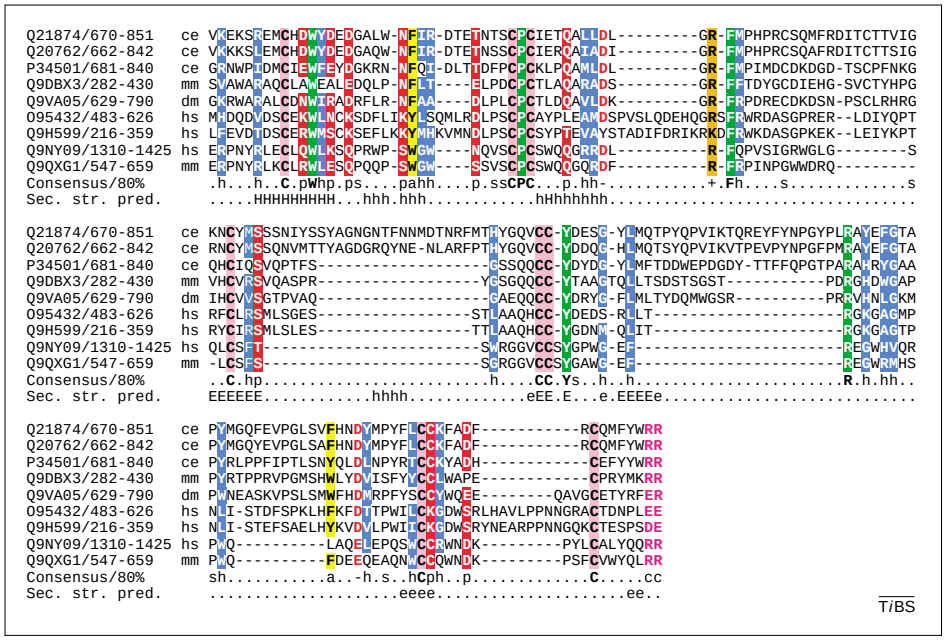

Fig. 2. Multiple sequence alignment of AMOP (adhesion-associated domain in MUC4 and other proteins) domains from different proteins; all domains were retrieved using PSI–BLAST [10]. In the first run of searching we retrieved hypothetical proteins from *Caenorhabditis elegans* (Q21874, Q20762, E = 10^{-22}) and mouse (Q9DBX3, E = 10^{-20}), and a protein involved in cell–cell adhesion from *Drosophila melanogaster* [19] (Q9VA05, E = 10^{-23}). The first PSI–BLAST iteration retrieved two other hypothetical proteins from human that show the same conserved cysteine pattern (O95432, E = 10^{-21} and Q9H599 E = 10^{-07}). The search converged at the second iteration. Using the alignment of all the sequences retrieved with PSI–BLAST to run a Hidden Markov Model search [20], we retrieved human and mouse MUC4 (E = 10^{-43} and E = 10^{-14}, respectively). Sequences are indicated using the database accession number followed by the starting and ending residues of the domain, and by the species. The consensus in 80% of the sequences is reported below the alignment: a, c, h, p, s, – and + indicate aromatic, charged, hydrophobic, polar, small, negative and positive residues, respectively. Hydrophobic residues are highlighted in blue, polar residues in red, aromatic residues in yellow, positive residues in orange, conserved cysteines in pink and other conserved residues in green. Negative residues are in red and charged residues in magenta. The secondary structure prediction (Sec. str. pred.) is taken from the consensus of the alignment (E, strand predicted with expected average accuracy >82%; e, strand predicted with expected average accuracy <82%) [11]. Abbreviations: ce, *Caenorhabditis elegans*; dm, *Drosophila melanogaster*; hs, *Homo sapiens*; mm, *Mus musculus*. This multiple sequence alignment (alignment number ALIGN_000277) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000277.dat).

noted that some proteins with a VWD domain show a different cysteine-rich domain, the VWC domain. Nevertheless, the number of conserved cysteines, as well as their positions, clearly differ between the VWC domain and our region.

AMOP is only found in proteins that also contain extracellular domains involved in cell adhesion (Fig. 1a), implying an extracellular localization. The alignment of all sequences with <80% identity that share this region shows that it is ~100 residues long and contains eight invariant cysteine residues (Fig. 2). The CYSPRED predictor [11], a neural network-based programme that predicts the bonding states of cysteines, suggests the involvement of these eight residues in disulfide bonds. This is another indicator of the extracellular localization of this domain. Secondary structure prediction using the PHD programme [12] indicates an initial region rich in α helix followed by four β strands, suggesting a β-sheet organization for the C terminus of the domain. However, despite using both the 3D-PSSM [13] and the THREADER [14] web servers, we failed to detect any significant similarity of our domain to a known protein fold or even to an already characterized pattern of disulfide bonds.

Functional implications

Recently, different splice variants of the full-length sv0-MUC4 protein have been identified in normal human testis, lung carcinoma [15] and different pancreatic tumour cell lines [16,17]. Some of these splice variants are soluble (e.g. sv1–sv7, sv9, sv11–sv19) whereas others are membrane-bound (e.g. sv0, sv8, sv10, sv20, sv21); their role in normal and cancer cells is unknown. Interestingly, the AMOP domain is only present in some of the splice variants characterized so far. Furthermore, although this domain is present in human and mouse full-length MUC4, it is absent from the potential rat orthologue. This suggests the existence of another possible splice variant and also supports the notion of AMOP as an independent protein module. So far, in addition to this rat orthologue, two MUC4 splice variants, sv6-MUC4 and sv8-MUC4 [15,18], lack AMOP domains (Fig. 1). Although the different function of these variants is not yet known, because of its presence in cell adhesion molecules we speculate that the AMOP domain is involved in adhesion processes. Furthermore, by exploitation of its extracellular localization, the AMOP domain could be used as an early prognostic marker for pancreatic carcinoma.

Acknowledgements

We are grateful to D. Torrents and M. Suyama for their help in the intergenic DNA analysis. The work is supported by NATO and DFG. F.D.C. is supported by an Istituto Mario Negri-Milano fellowship.

References

1 Guzman, K. *et al.* (1996) Quantitation of mucin RNA by PCR reveals induction of both MUC2 and MUC5AC mRNA levels by retinoids. *Am. J. Physiol.* 270, L846–L853
2 Braga, V.M. *et al.* (1992) Spatial and temporal expression of an epithelial mucin, Muc-1, during mouse development. *Development* 115, 427–437
3 Gendler, S.J. *et al.* (1995) Epithelial mucin genes. *Annu. Rev. Physiol.* 57, 607–634
4 Dekker, J. *et al.* (2002) The MUC family: an obituary. *Trends Biochem. Sci.* 27, 126–131
5 Ogata, S. *et al.* (1992) Mucin gene expression in colonic tissues and cell lines. *Cancer Res.* 52, 5971–5978
6 Balague, C. *et al.* (1994) Altered expression of MUC2, MUC4, and MUC5 mucin genes in pancreas tissues and cancer cell lines. *Gastroenterology* 106, 1054–1061
7 Balague, C. *et al.* (1995) *In situ* hybridization shows distinct patterns of mucin gene expression in normal, benign, and malignant pancreas tissues. *Gastroenterology* 109, 953–964
8 Schultz, J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5857–5864
9 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234
10 Altshul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
11 Fariselli, P. *et al.* (1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins* 36, 340–346
12 Rost, B. *et al.* (1994) PHD an automatic mail server for protein secondary structure prediction. *CABIOS* 10, 53–60
13 Kelley, L.A. *et al.* (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299, 499–520
14 Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold method for genomic

sequences. *J. Mol. Biol.* 287, 797–815
15 Moniaux, N. *et al.* (2000) Alternative splicing generates a family of putative secreted and membrane-associated MUC4 mucins. *Eur. J. Biochem.* 267, 4536–4544
16 Choundhury, A. *et al.* (2000) Human MUC4 mucin cDNA and its variants in pancreatic carcinoma. *J. Biochem.* 128, 233–243
17 Choundhury, A. *et al.* (2001) Alternate splicing at the 3′-end of the human pancreatic tumor-associated mucin MUC4 cDNA. *Teratog. Carcinog. Mutagen.* 21, 83–96
18 Moniaux, N. *et al.* (1999) Complete sequence of the human mucin MUC4: a putative cell membrane-associated mucin. *Biochem. J.* 338, 325–333
19 Hynes, R.O. *et al.* (2000) The evolution of cell adhesion. *J. Cell Biol.* 150, F89–F96
20 Eddy, S.R. (1998) Profile Hidden Markov models. *Bioinformatics* 14, 755–763

Francesca D. Ciccarelli*
Tobias Doerks
Peer Bork
European Molecular Biology Laboratory, 69012 Heidelberg, Meyerhofstr. 1, Max-Delbrueck-Centrum, PO Box 740238, D-13092 Berlin, Germany.
*e-mail: ciccarel@embl-heidelberg.de

A second catalytic domain in the Elp3 histone acetyltransferases: a candidate for histone demethylase activity?

Yurii Chinenov

A new subfamily of two-domain histone acetyltransferases (HATs) related to Elp3 has been identified. In addition to a HAT domain in the C terminus, these proteins have an N-terminal domain similar to the catalytic domain of S-adenosylmethionine radical enzymes. Two-domain organization is preserved in evolution, suggesting that both enzymatic activities are functionally or mechanistically coupled and directed towards highly conserved substrates. The functional implications of this similarity and a possible role for Elp3-related proteins as histone demethylases are discussed.

Histone N-terminal modifications affect gene expression by altering the repertoire of chromatin-associated proteins and, consequently, the compaction state of the chromatin. Methylation of specific residues in histones correlates with the persistently repressed state of heterochromatin [1–4]. The relative stability of this modification led to the hypothesis that certain methyl groups in histones provide a long-lasting epigenetic mark committing the chromatin to a specific transcriptional state [5,6]. However, recently reported transient repression of euchromatic genes associated with histone methylation suggests that histone methylation is less permanent than previously thought and that mechanisms of regulated removal of methyl groups must exist [7,8]. Several mechanisms, including specific degradation of methylated histones or enzymatic demethylation, have been proposed. The histone demethylase activity was described almost 30 years ago [9], but the protein that is responsible for this activity has not yet been identified. Here, I report the sequence similarities of the yeast histone-acetyltransferase (HAT) Elp3 to an enzyme superfamily that uses S-adenosylmethionine (SAM) in radical reactions. I also discuss the possible role of Elp3 as a histone demethylase.

The similarity between Elp3-related HATs and SAM radical enzymes

Screening for proteins similar to oxidases involved in heme biosynthesis revealed an extremely well-conserved group related to Elp3 [10]. All members of this group contain the C-terminal HAT domain and a separate, previously unreported region similar to the catalytic domain of both bacterial anaerobic coproporphyrinogen III oxidases (COPIII) and other proteins belonging to the SAM radical family. This similarity implies that, in addition to HAT activity, Elp3-related proteins might possess a second enzymatic activity. The conserved two-domain organization of Elp3-related proteins (Figs 1,2a) and high degree of sequence identity (81% in COPIII domain between human and yeast compared with 64% for cytochrome *c* and 81% for histone H2B) suggest that both enzymatic activities are functionally or mechanistically coupled and that the substrate or substrates of these enzymes are also well conserved. As two-domain Elp3-related proteins have been found in both *Eukaryota* and *Archaea* but not in *Bacteria*, the conserved substrates of Elp3 enzymatic activities are probably specific for these two kingdoms and absent from *Bacteria*.

Anaerobic COPIII belongs to the SAM radical enzyme family, members of which use SAM in a wide range of reactions including oxidation, oxidative cyclization, N- and P-methylation, isomerization and protein radical formation [11,12].
A conserved glycine-rich region in both Elp3 and SAM radical enzymes (Fig. 1, marked by a blue bar) is similar to the motif 1 in several SAM-dependent methyltransferases [13] and might be involved in SAM binding. Both Elp3-related proteins and SAM radical enzymes contain several conserved acidic and glycine residues reminiscent of motifs II and III of SAM-dependent methyltransferases.

Coproporphyrinogen III oxidase catalyses the removal of the carboxyl group and two hydrogens of propionic groups in coproporphyrinogen III to form vinyl groups of protoporphyrinogen IX during haem biosynthesis [14]. In the context of chromatin, the COPIII-related domain of Elp3 could catalyse the modification of DNA, histones or other chromatin-associated proteins. Similarity to COPIII suggests that Elp3-related proteins catalyse a reaction involving elimination of a single-carbon fragment. Thus, it is possible that Elp3-related proteins are involved in demethylation of DNA or histones. As Elp3-related proteins have been found in organisms in which DNA methylation has not been detected (e.g. yeast, *Caenorhabditis elegans*), methylated DNA is a less likely substrate.

Putative catalytic mechanism for histone demethylation

Similar to SAM radical enzymes, a histone demethylation reaction might be initiated by scission of SAM into methionine and 5′-deoxyadenosyl radical that is involved in further catalytic steps. In SAM radical

  PII: S0968-0004(02)02058-3